The Korean Journal of Applied Statistics (2008)
21(4), 659–664

Application of Multiple Imputation Method in Analyzing Data with Missing Continuous Covariates

S. Ghasemizadeh Tamar[1] · M. Ganjali[2]

[1]Dept. of Statistics, Shahid Beheshti University; [2]Dept. of Statistics, Shahid Beheshti University

(Received March 2008; accepted June 2008)

Abstract

Missing continuous covariates are pervasive in the use of generalized linear models for medical data. Multiple imputation is the most common and easy-to-do method of dealing with missing covariate data. However, there are always serious warnings in using this method. There should be concern to make imputed values more proper. In this paper, proper imputation from posterior predictive distribution is developed for implementing with arbitrary priors. We use empirical distribution of the posterior for approximating the posterior predictive distribution, to sample from it. This method is preferable in comparison with a presented imputation method of us which uses a full model to impute missing values using available software. The proposed methods are implemented on glucocorticoid data.

Keywords: Generalized linear model, missing data, proper multiple imputation, predictive distribution, full imputation model.

1. Introduction

Missing covariate data in generalized linear models is a common issue in clinical trials and observational studies. We consider a regression analysis of a response variable y on a vector of covariates, z, that are fully observed and a continuous covariates, x, that is unobserved for some subjects. Such analysis, with existing software packages, requires full covariate information, so a simple way to overcome the missing data problem is to analyze just completely observed subjects. This method is known as complete case(CC) analysis. The CC analysis can be biased, if the data are not missing completely at random(MCAR, this means the failure to observe a value does not depend on any data, either observed or missing, see Rubin, 1976). When the failure to observe a value depends on the value that could have been observed, the data are said to be not missing at random(NMAR). If the failure to observe a value, conditional on observed data, dose not depend on the data that are unobserved, the data are said to be missing at random(MAR). Even if the CC analysis of our data is unbiased, the efficiency is decreased by increasing the fraction of missing values. Thus it is reasonable to seek new methods of handling missing data, appropriately.

[1]Professor, Dept. of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran. E-mail: s.tamar@mail.sbu.ac.ir

[2]Corresponding author: Professor, Dept. of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran. E-mail: m-ganjali@sbu.ac.ir

In this paper we examine multiple imputation(MI) for inference in generalized linear models(GLMs). The MI has been viewed as a general answer to the missing data problem in recent years. Multiple imputation is the simplest and widely used method, in which specific values are used to fill in the missing data, introduced by Rubin (1977a, 1977b, 1987). This technique involves creating some copy of the dataset and filling in the missing data with specific values. Then each completed datasets is analyzed separately and the results combined into one result by averaging over the completed datasets (Little and Rubin, 2002). Rubin (1987), Little and Rubin (2002) and van Buuren and Oudshoorn (1999) have discussed MI method for missing covariates in GLMs.

In this paper, assuming to have MAR data, we present a simple imputation method which can be implemented by existing software such as SAS. This simple method uses a full model to impute the missing continuous covariates. Multiple imputation will be used to find parameter estimates of the proposed method of imputation. Also we shall use the empirical predictive distribution to generate proper values of multiple imputation for missing continuous covariates. It makes the method proposed by Ibrahim *et al.* (2005) available for arbitrary priors. The two methods will be also compared.

The rest of this paper is organized as follows: In Section 2, to have a motivation, we present the data set. In Section 3, we introduce briefly multiple imputation method, empirical posterior predictive distribution and our simple method for imputation. In Section 4, the results of these methods on the data set are given and compared. We have a discussion in Section 5.

2. Motivation

There is a substantial clinical and laboratory evidence of the efficacy of glucocorticoid in the treatment of acute pulmonary surfactant deficiency in preterm newborns. The hypothesis that maternal antenatal glucocorticoid receipt is followed by reduced risk of bronchopulmonary dysplasia(BPD) had been explored by van Marter *et al.* (1990). A sample of 223 intubated infants weighing less than 1,751gm birth weight provided 76 infants with BPD(defined by both oxygen requirement and compatible chest radiograph) and 147 who had not BPD characteristic by day 28 of life. The response y (BPD) variable is binary and indicates having BPD or not, by the day 28 of life. About 34% of infants had BPD. Birth weight(denoted by z) was recorded in 100gm. All infants were preterm newborns and the age(denoted by x) was recorded in weeks. We have the full data set in this application, but for comparative purposes we create missing values in covariate age by a MAR mechanism. We choose age for this aim, since the age records are subject to missingness, specially in developing countries, where women are not under medical surveillance during their pregnancy, so recording is not exact and may be assumed to be missed. For 40% of subjects the age variable's values are missed under MAR mechanism *i.e.* missingness depends on observed variables and not on values of age. To do so, the vector parameter $\gamma = (\gamma_1, \gamma_2)$ in $P(R_i = 1 \mid y_i) = \Phi(\gamma_1 + \gamma_2 y_i)$ is chosen such that the age variable is observed for nearly 80% of healthy infants and nearly 65% of the rest($\gamma_1 = 0.84$, $\gamma_2 = -0.46$).

Table 2.1 shows the result of using a simple logistic model (See, model (2.1) in Subsection 2.1) for full data and after removing cases with missing x (CC analysis). As can be seen, the CC analysis overestimates the age and weight effects, but underestimates the intercept parameter. The variance of parameter estimates are also overestimated by CC analysis. For full data, age and weight both have negative effects on the odds of BPD. This means that the older the child the less likely she/he has BPD and also the heavier the child the less likely she/he has BPD.

Table 2.1. Results of using logistic model for full data and complete case(CC)

Parameter	Full		CC	
	Est.	S.E.	Est.	S.E.
intercept	13.83	2.930	12.82	3.450
age	−0.39	0.113	−0.38	0.132
weight	−0.24	0.080	−0.21	0.090

2.1. Models for response and missing covariate

As some of the x values(age values) are missing we have to define a distribution function for x. We consider a logistic model for the binary response and a normal regression model for covariates($\mathbf{x}$) as below:

$$\text{logit}\,[P(y_i = 1 \mid x,\, z)] = \beta_0 + \beta_1 x_i + \beta_2 z_i \tag{2.1}$$

and

$$X_i \mid z_i = \alpha_0 + \alpha_1 z_i + \epsilon_i, \quad \epsilon_i \sim N(0,\, \sigma^2), \quad i = 1, \ldots, n, \tag{2.2}$$

where $\text{logit}(a) = \ln[a/(1-a)]$.
So we decompose the joint distribution of y and x given z as a conditional distribution for y given x and z and a conditional distribution for x given z.

3. Methods

3.1. Multiple imputation

Imputation is a common and flexible way to fill incomplete data. In this method, each missing value in the data set is filling with a drawout from a predictive distribution and the resulting complete data set is analyzed via standard techniques. However there is a serious lack of taking into account the variation due to not considering uncertainty of imputed values. Multiple imputation concerns this uncertainty by iterating the imputed data. The steps of multiple imputation are as below:

- Create M copies of data sets by imputation techniques.
- Analyze the data sets and estimate vector parameter $\gamma^{(m)}$ for the m^{th} imputed data set, $m = 1, \ldots, M$.
- Combine the results and get the MI parameters, *i.e*:

$$\hat{\gamma}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\gamma}^{(m)},$$

$$\hat{V}_{MI} = \bar{V} + \left(1 + \frac{1}{M}\right) \hat{B},$$

where $\hat{B} = 1/(M-1) \sum_{m=1}^{M} (\hat{\gamma}^{(m)} - \hat{\gamma}_{MI})(\hat{\gamma}^{(m)} - \hat{\gamma}_{MI})'$ and $\bar{V} = 1/M \sum_{m=1}^{M} \hat{V}^{(m)}$, average within imputation variance-covariance matrix.

We use $M = 5$ in this paper, which is good enough to give efficient estimates (See, Rubin, 1987). The first step in doing multiple imputation is to make drawout from the predictive distribution, in our example $f(x \mid y, z, \gamma)$. In the following we first present our simple method, then we remind use of the proper predictive distribution and we extend it to use of empirical posterior predictive distribution.

3.2. The simple method of imputation

Let $f_0(x \mid z, \gamma) = f(x \mid y = 0, z, \gamma)$ and $f_1(x \mid z, \gamma) = f(x \mid y = 1, z, \gamma)$. Since the distribution of (x, z) is Multivariate normal, it is suitable to use MI procedure of SAS software to compute imputation values. Notice that the model which we use for imputing the missing values of variable x is:

$$\begin{aligned} E(X \mid y, z) &= \alpha_{0y} + \alpha_{1y} z \\ &= \begin{cases} \alpha_{00} + \alpha_{10} z, & y = 0, \\ \alpha_{01} + \alpha_{11} z, & y = 1, \end{cases} \end{aligned} \tag{3.1}$$

which may be written as below:

$$E(X \mid y, z) = \alpha_0^* + \alpha_1^* y + \alpha_2^* z + \alpha_3^* yz,$$

where $\alpha_0^* = \alpha_{00}$, $\alpha_0^* + \alpha_1^* = \alpha_{01}$, $\alpha_2^* = \alpha_{10}$ and $\alpha_2^* + \alpha_3^* = \alpha_{11}$, *i.e.* the model includes interaction effects of y and z. This imputation model involves all available information and is expected to give good imputed values. Note that the data set is divided into two parts (conditioned on y) and each part is imputed separately and finally merged in one data set, before analysis. The MI procedure of SAS software uses Gibbs sampler to generate sample from $f_j(x, \alpha_j, \sigma_j \mid z)$ for $j = 0, 1$ and $\alpha_j = (\alpha_{0j}, \alpha_{1j})'$. The steps are as below:

- Assuming k^{th} iteration, generate a value for x_i from conditional distribution of x_i on z_i, *i.e.* $f_j(x_i \mid z_i, \alpha_j^{(k)}, \sigma_j^{(k)})$.
- Generate $(\alpha_j^{(k+1)}, \sigma_j^{(k+1)})$ from $f_j(\alpha_j, \sigma_j \mid x_i, z_i)$, $j = 0, 1$. These new estimates are then used in the first step. Here we use a noninformative Jeffrey's prior.

The two steps are iterated long enough for the results to be reliable for a multiply imputed data set (Schafer, 1997, p.72). The goal is to have the iterates converge to their stationary distribution and then to simulate an approximately independent draw of the missing values.

3.3. The use of proper predictive distribution for imputation

The second method of imputation is to use the proper predictive distribution of $(x \mid y, z)$, *i.e.*:

$$P(x_i \mid y_i, z_i) \propto \int\int P(x_i \mid y_i, z_i, \beta, \alpha) \pi(\beta, \alpha \mid y, z) d\beta d\alpha, \tag{3.2}$$

where $\pi(\beta, \alpha \mid y, z)$ is the joint posterior distribution for (β, α) based on observed data (*cf.* Ibrahim *et al.*, 2005). Typically in practice, a noninformative prior such as a uniform improper prior, or a joint normal prior with noninformative choices for hyperparameters is chosen for (β, α). Sampling from $P(x_i \mid y_i, z_i)$ is not an easy work for arbitrary priors and different densities of $(x \mid y, z, \alpha)$ (Indeed it has been done just for the multivariate normal distributions and improper uniform prior, Ibrahim *et al.*, 2005).

We try to approximate predictive posterior distribution, $P(x_i \mid y_i, z_i)$, at first, by empirical posterior distribution and then sample from it. Empirical distribution of $\pi(\beta, \alpha \mid y, z)$ can be computed by simulating sample via the method proposed by Ibrahim *et al.* (2002):

- Specify $P(x \mid y, z, \alpha)$.
- Specify the prior, $\pi(\beta, \alpha)$.
- To sample from $P(\beta, \alpha \mid y, z)$, do the following steps:

Table 4.1. Resuts of MI from full model and empirical posterior predictive distribution

Parameter	Full model		Empirical posterior	
	Est.	S.E.	Est.	S.E.
intercept	14.38	3.34	13.80	3.44
age	−0.41	0.16	−0.39	0.13
weight	−0.24	0.08	−0.25	0.08

- Sample β from $P(\beta \mid y, z, \alpha, x)$
- Sample α from $P(\alpha \mid y, z, \beta, x)$
- Sample x from $P(x \mid y, z, \beta, \alpha)$

- Use the sample to compute the empirical posterior distribution,

$$\pi^*((\beta, \alpha) \in h) = w_h = \frac{\sum_{k=1}^{N} I_{\{h\}}\left(\beta_{(k)},\, \alpha_{(k)}\right)}{N}, \tag{3.3}$$

  where $I_{\{h\}}(\beta,\, \alpha) = 1$ if $(\beta,\, \alpha)$ is belong to the h^{th} rectangle and it is 0 otherwise.

- Substitute for the empirical posterior distribution in (3.2) as below

$$p(x_i \mid y_i,\, z_i) \propto \sum_{h=1}^{K} p\left(x_i \mid y_i,\, z_i, \beta_{(h)},\, \alpha_{(h)}\right) w_h, \tag{3.4}$$

  where K is the total number of rectangles and $\beta_{(h)}$ and $\alpha_{(h)}$ are the avarege of points in h^{th} rectangle.

- Generate imputations from (3.4).

Generating imputations is simple via the acceptance sampling. The advantage of this proper imputation is that it is not restricted to the specific prior or distributions.

4. Results

Table 4.1 shows the results of using the multiple imputation for two imputation methods, imputation from the full model (3.1) and imputation from the empirical posterior predictive distribution(EPPD) (3.4). For implementing EPPD, firstly we simulate $\pi(\beta, \alpha \mid y, z)$ with a sample of size 20000 and then we use acceptance sampling to generate imputed values. Here this method is preferable to the full model imputation of (3.1) since the results of using this method are closer to the results of usnig full data in Table 2.1. Since the efficient full models for imputation are usually the best, if available (See Nielsen, 2003), we use full model here to compare our proper imputaton with empirical posterior predictive distribution with a usually best method. However, method of full model is good enough in comparing with CC analysis.

5. Discussion

We have presented a simple multiple imputation such that one can use the multivariate normal imputation method of SAS software to analyze data with missing in covariate with binary response. The method can be extended to be used for nominal responses in a manner similar to what we did

for binary response. Also we have developed use of posterior predictive distribution, in such a way that arbitrary densities and different priors, for generating x from $f(x \mid y, z)$ can be implemented. Our example shows that both method performs well, but the simple method is better to be used due to its precision, simplicity and ability of using existing software. The simple method discussed here can be used for data with a vector of missing continuous covariate. For this, one needs to assume multivariate normal distribution for x in multiple imputation approach.

Acknowledgements

The correspondence author would like to thank Shahid Beheshti University for the awarded grant.

References

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **30**, 55–78.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review, *Journal of the American Statistical Association*, **100**, 332–347.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Nielsen, S. F. (2003). Proper and improper multiple imputation, *International Statistical Review*, **71**, 593–607.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin, D. B. (1977a). Formalizing subjective notions about the effect of nonrespondents in sample surveys, *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D. B. (1977b). *The Design of a General and Flexible System for Handling Non-Response in Sample Surveys*, working document prepared for the U.S. Social Security Administration.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, New York.

van Buuren, S. and Oudshoorn, K. (1999). Flexible multivariate imputation by MICE, *Leiden, The Netherlands: TNO Prevention Center.*

van Marter, L. J., Leviton, A., Kuban, K. C. K., Pagano, M. and Allred, E. N. (1990). Maternal glucocorticoid therapy and reduced risk of bronchopulmonary dysplasia, *Pediatrics*, **86**, 331–336.